Complementing real datasets with simulated data: a regression-based approach

Complementing Real Datasets with Simulated Data: A Regression-based Approach

M. A. Ortiz-Barrios[*], J. Lundström[†], J. Synnott[‡], E. Järpe and A. Sant'Anna[§]

March 6, 2020

Abstract

Activity recognition in smart environments is essential for ensuring the wellbeing of older residents. By tracking activities of daily living (ADLs), a person's health status can be monitored over time. Nonetheless, accurate activity classification must overcome the fact that each person performs ADLs in different ways and in homes with different layouts. One possible solution is to obtain large amounts of data to train a supervised classifier. Data collection in real environments, however, is very expensive and cannot contain every possible variation of how different ADLs are performed. A more cost-effective solution is to generate a variety of simulated scenarios and synthesize large amounts of data. ~~The challenge then becomes ensuring that simulated data is a reliable representation of real data~~. Nonetheless, simulated data can be considerably different from real data. Therefore, this paper proposes the use of regression models to better approximate real observations based on simulated data. ~~This paper compares~~ To achieve this, ADL data from a smart home were first compared with equivalent ADLs performed in a simulator. ~~The statistical analysis is based on~~ Such comparison was undertaken considering the number of events per activity, number of events per type of sensor per activity, and activity duration. Then, ~~we assessed~~ different regression models were assessed for calculating real data based on simulated data. The results evidenced that simulated data can be transformed with a prediction accuracy $R^2 = 97.03\%$.

Keywords: Activity recognition, Activity duration, Regression analysis, Non-linear models, Determination coefficient, Quantile-quantile plots

1 Introduction

The global population is ageing due to improvements in public health, increased life expectancy, and falling fertility rates.[1] The number of people aged 60 years or older

[*]Department of Industrial Management, Agroindustry and Operations, Universidad de la Costa CUC, Barranquilla, Colombia, mortiz1@cuc.edu.co

[†]Convergia Consulting, Halmstad, Sweden, jens@convergia-consulting.io

[‡]School of Computing, Computer Science Research Institute, Ulster University, Belfast, BT37 0QB, UK, j.synnott@ulster.ac.uk

[§]Department of Intelligent Systems and Digital Design, Halmstad University, Halmstad, Sweden, eric.jarpe@hh.se and anita.santanna@hh.se

worldwide is projected to grow from 0.9 billion to 1.4 billion between 2015 and 2050. Within this age range, the fastest growth is anticipated in those aged 80 or over, with estimates indicating increases from 125 million in 2015 to 434 million by 2050.[2]

Older adults are afflicted by 23.1% of the global burden of disease. This is 49.2% of the burden in high-income regions, and 19.9% of the burden in low and middle-income regions. The most burdensome health issues include ischaemic heart disease, stroke, diabetes, falls, dementia and depression.[3] The ageing population has seen an increase in the prevalence of such conditions. For example, as of 2018, dementia affects 50 million people worldwide. This is predicted to increase to 152 million by 2050.[4]

Despite these chronic health conditions, it is ~~not un~~common for older adults to live alone. Recent reports indicate that 26% (12.1 million) of older adults in the United States, and 32% (3.65 million) of older adults in the UK live alone.[5,6] Care for chronic disease often requires long-term close monitoring, which is resource intensive. There are indications that health systems around the world are struggling to cope with this increasing demand.[7] For example, an investigation into the UK domiciliary care market suggested that publicly funded access to domiciliary care is been reduced and restricted to those with the greatest needs due to budget constraints.[8]

There is therefore the need for innovative, technology-based approaches to to help alleviate the strain of these increasing demands on increasingly limited healthcare resources. Such technology-based approaches may be used to improve the cost-effectiveness of domiciliary care through data-driven decision making and more efficient use of resources. In addition to maximizing the coverage of care service offerings, these approaches also aim to produce increased quality of care through objective rather than subjective decision making, early detection of conditions, prediction of change in condition, and more detailed and earlier insight into the impact of intervention.[9] One area of interest is the automatic analysis of activities of daily living (ADLs) performed by older adults living alone. ADLs consist of a range of activities that are required to manage basic physical needs. These activities span areas including grooming and personal hygiene, dressing, toileting and continence, ambulation, and eating.[10] Independent performance of ADLs is correlated with physical and cognitive function. Increased dependency in performing ADLs has been associated with dementia,[10] hospitalisation, morbidity and mortality.[11] Previous works have suggested that ADL monitoring may facilitate the early detection of conditions such as dementia.[12]

Activity recognition is the process of automatically recording, identifying and analysing the performance of activities by processing sensor data.[13] Sensors typically deployed in the home environment include door contact sensors, passive infrared (PIR) sensors, pressure sensors, audio sensors, accelerometers, thermal sensors. Activity recognition depends on the creation of accurate and generalizable classification models.[14,15] The creation of such models relies upon the availability of realistic activity data.[16] However, compiling high quality, large datasets is difficult due to large costs, lack of flexibility and scalability of intelligent environment construction, as well as the practical limitations of recording a comprehensive range of activities with all possible variations.[17,18]

One approach to overcome learning ADLs from a limited dataset collected in the wild is to use *transfer learning*. By adopting machine learning models to capture the intrinsic properties of human behaviour (e.g. ADLs) in one home, it is hypothesized

that the model can be re-used to augment the learning of another person's ADLs in another home. A particular challenge here is to make the process unsupervised.[19]

The barriers to the collection and availability of activity data have been said to be detrimental to research progress and may slow advances in the field.[20,21] Researchers have been exploring the application of simulation approaches to generate synthetic activity datasets. Simulation can provide a mechanism of rapidly generating vast datasets spanning extended periods of time without the need for investment in physical systems nor recruitment of research subjects. Synnott et al[17] provide a comprehensive overview of approaches to the simulation of smart home activity data. Recently, Alshammari et al[14] have developed the OpenSHS smart home simulator, which facilitates data generation through a hybrid approach combining both interactive and model-based approaches. This simulator has been used to produce datasets for classification and anomaly detection.[22] Francillette et al[23] have developed an intelligent environment simulator capable of generating data from simulated sensors such as RFID, ultrasound, pressure sensors, and contact sensors, amongst others. Lee et al[24] developed the Persim 3D human activity simulator. Kamara-Esteban et al[25] created MASSHA, which is an agent-based simulator for simulating activities within intelligent environments. Synnott et al[26] created IE Sim, an intelligent environment simulation tool that has previously been used to generate a benchmark dataset shared by three international research organisations.

The existing approaches to the simulation of ADL data have provided excellent steps towards producing synthetic data for experimentation and development of novel approaches. Nevertheless, an ongoing challenge in developing such simulation software is the ability to generate simulated data that accurately represents real data.[24] A comparison between real data collected within the Gator Tech Smart House and simulated data generated by Persim 3D[24] revealed average data similarities of between 78% and 81%. Another study comparing real data ~~and~~ with data generated using the simulator MASSHA[25] found similarity to be between 88.10% and 93.52% in terms of frequency, and 98.27% and 99.09% in terms of duration on datasets containing single user activities.

In light of the reported literature, the evidence base directly concentrating on comparing real observations and simulated data is largely limited and poorly developed. ~~In order to improve the similarity between real and simulated data~~ To address this gap in knowledge, it is important to consider which activity is being performed (i.e. which sensors are being triggered) as well as the duration and timing of these activities (i.e. the duration and intensity of sensor triggers). For example, preparing a meal might have a longer duration in the evening than in the morning.

Existing interactive approaches that rely primarily on avatar and simulated environment interaction[17] to generate synthetic datasets have a limited ability to take into account the natural differences in activity duration and intensity in relation to time of day. This is primarily due to the artificial nature by which interaction takes place. As a result, dataset similarities will not be optimal. To improve upon this, we propose the application of regression modeling in order to capture activity duration and intensity given the time of day. We show that our linear and non-linear models can improve the similarity of activity data from different environments. Consequently, the contribution of this paper will be two-fold: i)Verification of the similarity between real and simulated data in terms of activity duration, number of events per activity, and number of events per type of sensor per activity, ii)

Proposal of models that better approximate real observations using data provided by the simulators.

The remaining Section 2 of this paper presents the methods and models used within the study. Results are presented and discussed in Section 3 and conclusions made in Section 4.

2 Method

This section briefly presents the models (linear, logarithmic, quadratic and square root) and hypothesis tests (ANOVA, Durbin-Watson and Anderson-Darling) used in this work. More thorough presentations of these methods can be found in various textbooks. We refer the reader to [27] for a basic introduction and[28] and[29] for more extensive introductions.

The most common regression model is the *linear* (or *additive*) *model*, $Y = \beta_0 + \boldsymbol{\beta}\boldsymbol{X} + \epsilon$. Here Y is called *response*, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$) are *covariates*, and ϵ is the *residual*. When there is no linear dependence between variables, the model can be simplified as:

$$Y = \beta_0 + \sum_{k=1}^{n} \beta_k X_k + \epsilon \tag{1}$$

In the simplified case, the parameters of the model are the *intercept* β_0 and *regression coefficients* $\beta_1, \beta_2, \ldots, \beta_n$ (and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$). This model is also referred to as a *linear regression*. The model in Eq. 1 is the most commonly used. It makes the least the least assumptions about dependence mechanisms, so it is frequently the choice where the application does not motivate interdependent covariates.

In order to assess whether a linear model is appropriate, one may test whether the regression coefficients are non-zero to justify inclusion of the variables in the model. In addition, one should calculate the determination coefficients to assess the model fit, and make Quantile-Quantile plots to check that the assumption residuals are normally distributed, see subsection 2.2.

A means to sort data into different categories may be achieved by using *dummy variables*.[30] These are binary variables which are 1 for all data where the individual satisfy some criteria, and 0 for all data when the individual does not.

2.1 Transforms of the linear model

Of course variables may depend on each other in a non-linear way. In which case, a carefully investigated linear model can be transformed so as to capture more complex dependencies between variables.

Provided that the response variable is positive, the *logarithmic* (or *multiplicative*) *transform*

$$Y = \beta_0 \left(\prod_{k=1}^{n} \exp(X_k)^{\beta_k} \right) \exp(\epsilon) \tag{2}$$

is equivalent to saying that $\log Y$ is linearly dependent with X. By logging both sides, the model in Eq. 2 may alternatively be expressed

$$\log Y = \beta_0' + \sum_{k=1}^{n} \beta_k X_k + \epsilon \tag{3}$$

where $\beta_0' = \log \beta_1$.

A logarithmic model according the Eq. 2 and 3 is the right one when the logged response is proportional to a linear combination of the logged covariates. An example is population growth: if the response is the population size, this can be modelled as an initial size to some power. The exponent may be interpreted as the number of the generation. Covariates may be time elapsed from start and environmental variables.

In modeling activity duration, Y, by means of time of opening the bedroom door, X_1 and pressure values in a bed sensor, X_2, it may be that the relationship is neither well modeled by a linear nor a logarithmic model. If increased levels of X_1 or X_2 adds more to Y than would have been the case with just a linear model, it may be that a model which includes quadratic covariates, X_1^2 or X_2^2, is justified.

The *quadratic transform* is defined by

$$Y = \beta_0 + \sum_{k=1}^{n} \left(\beta_{1,k} X_k + \beta_{2,k} X_k^2 \right) + \epsilon. \tag{4}$$

If the non-linear effect is even stronger, a higher degree polynomial could be motivated. For a non-negative response variable a *squared quadratic transform*

$$\sqrt{Y} = \beta_0 + \sum_{k=1}^{n} \left(\beta_{1,k} X_k + \beta_{2,k} X_k^2 \right) + \epsilon \tag{5}$$

may be defined as given in the given equation. Both the quadratic transform in Eq. 4 and squared quadratic transform in Eq. 5 may be referred to as polynomial regressions.[31]

2.2 Assessment of the regression model

The validity of any model should be checked by testing that the regression coefficients are non-zero, and ANOVA may be used to this end.

To justify the inclusion of variables in the model, correlation analysis may be used to check which variables are highly correlated with the response or with the residuals from a previous model.

The residuals of a regression model are assumed be independent of each other so the the autocorrelation of residuals should be zero, a property which may be measured by Eq. 6. The Durbin-Watson statistic

$$DW = \frac{\sum_{i=2}^{N} (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n} \epsilon_i^2} \tag{6}$$

can be used to estimate these autocorrelations.

In order to check the assumption of normally distributed residuals, quantiles of the observed sample could be plotted against corresponding quantiles of the normal distribution. This is often referred to as *Quantile-Quantile plots* or *QQ-plots*. Also, an Anderson-Darling hypothesis test may be performed to check for deviation from normality of the residuals.

2.3 Experiment description

An experiment was conducted at the Halmstad Intelligent Home (HINT), Halmstad University, Sweden.[32] HINT is equipped with over 60 sensors including door contact sensors, ~~passive Infrared (~~PIR~~)~~ sensors, and pressure sensors, amongst others. The environment was designed to facilitate physiological monitoring, safety monitoring, functional monitoring, and emergency detection and response.[32] The left side of Fig. 1 shows HINT's floor plan. This floor plan was used to create a virtual environment within IE Sim, complete with virtual sensors (shown on the right side of Fig. 1).

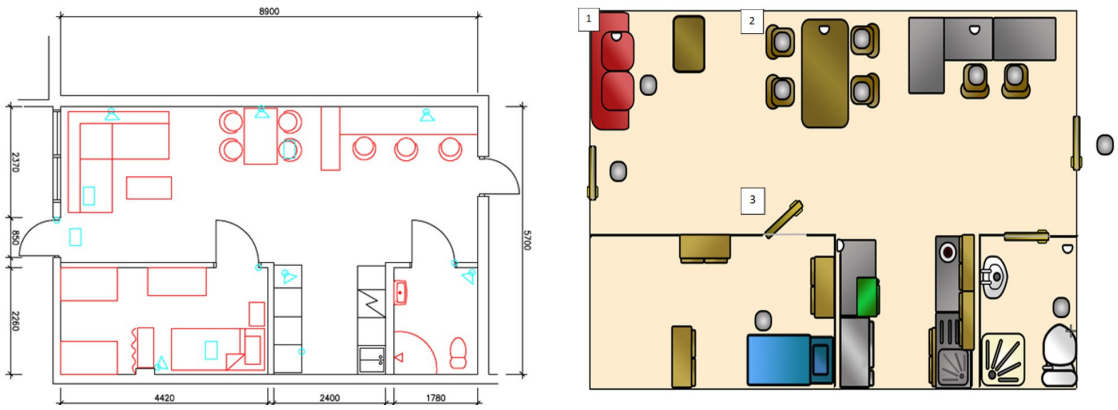

Figure 1: Left picture: The floor plan of the Halmstad Intelligent Home. Right picture: The virtual environment created within IE Sim, representing the Halmstad Intelligent Home. (1) PIR sensor; (2) Pressure sensor; (3) Door sensor.

Eleven participants were asked to perform a set of activities in the virtual environment by controlling a virtual avatar. Fig. 2 is the activity list that was provided to participants. Once participants had performed the activities within IE Sim, they were asked to perform the same activities (in person) within HINT. The output from the two data collections is structured as a list of events, where each event has a time stamp, sensor ID, sensor type (e.g. PIR or door sensor) and sensor state (e.g. open or closed). Moreover, at HINT the participants carried a button which was pressed when switching activity in order to ease annotation of the data set.

3 Results and Discussion

All 11 participants were able to complete the assigned tasks successfully. In total, 1105 simulated sensor events were generated, with a mean of 100.45 (SD: 29.97) sensor events per participant. The mean time taken per participant to complete all simulated activities was 521.45 (SD: 123.20) seconds. The participants then performed these activities at (HINT). As a result, 930 real sensor events were produced with an average of 116.25 (SD: 14.39) events per participant. The average time spent per each person to finish all the activities was 835.9 (SD: 213.42) seconds. Our analysis consisted of assessing how similar simulated data are to real data. Then, regression analysis was used to determine whether the simulated data could be used to predict real data.

ACTIVITY LIST
Note:
Please close each door after passing through
Please switch off each appliance after use
You will be guided through each activity in sequence, please remember to select the "Stop_Start_" buttom after each activity is complete
Time is not an issue with this experiment - do not worry about needing to take time to re-read an activity description, etc.
ACTIVITY 1 - GO TO BED
You can stay in bed all time that you want. Time maximum is 2 minutes. Then you have to leave bedroom, close the door and press the button
ACTIVITY 2 - USE BATHROOM
You can use toilet if you need, or just wash hands. Then leave bathroom, close door and press the button.
ACTIVITY 3 - PREPARE BREAKFAST
You have to prepare something to eat for breakfast. You can choose between milk & cereals or coffee, but you can make also prepare both. Then put the bowl on the table, sitting and press the button.
ACTIVITY 4 - LEAVE HOUSE
You can choose to leave the home either from the front door or from the garden door. When you are outside press the button.
ACTIVITY 5 - GET COLD DRINK
You can choose between tap water or by taking something from the fridge. Then put the glass with drink on the kitchen desk and push the botton.
ACTIVITY 6 - OFFICE
You have to go to office and write in the paper "Thanks for participating" and then press the button.
ACTIVITY 7 - GET HOT DRINK
You can choose between making tea or coffee. Then put the cup on the kitchen desk and press the button.
ACTIVITY 8 - PREPARE DINNER
You have to prepare a soup. Put the bowl on the table and press the button.

Figure 2: The activity list provided to participants.

3.1 Comparison of simulated data and real data

A paired t-test ($\alpha = 0.05, CL = 0.95$) was conducted to compare the simulated and real data in terms of the activity duration,[33] number of events per activity and number of events per type of sensor per activity. In this analysis, eight ADLs were considered: *Go to bed*, *Use bathroom*, *Prepare breakfast*, *Leave house*, *Get cold drink*, *Be in the office*, *Get hot drink*, and *Prepare dinner*.

Variable 1: *Activity duration* (AD)

Table 1 and Fig. 3 present the results of the comparative analysis between simulated and real activity duration for User 1 as an example. Given that the 95% confidence interval for the difference between the two activity duration values exclude zero; then the variables are statistically different. The p-value ($p = 0.013$) indicates that the data do not provide support for the null hypothesis, that is, the activity duration derived from simulated and real-environment are not statistically equivalent in User 1. Specifically, activity duration from real-environment ($\mu = 137.1s$) is significantly higher than activity duration from simulation ($\mu = 46.5s$). Therefore, the null hypothesis is rejected and we conclude, in the case of User 1, that there statistically significant difference between simulation and real-environment in terms of activity duration with a confidence level of 95%.

A summary of the results from the comparative analysis for all users can be found in Table 3. Based on a paired t-tests for activity duration, we found that in 75% of the users, the null hypothesis was rejected (p-value < 0.05). Therefore, it can be concluded that the activity duration derived from the simulated and real environments tend to be statistically different with a confidence level of 95%.

The next step is to identify the causes of this difference. Future work should determine whether differences are more common during performance of specific ADLs. Additionally, it is also recommended to study the profile of the users who performed

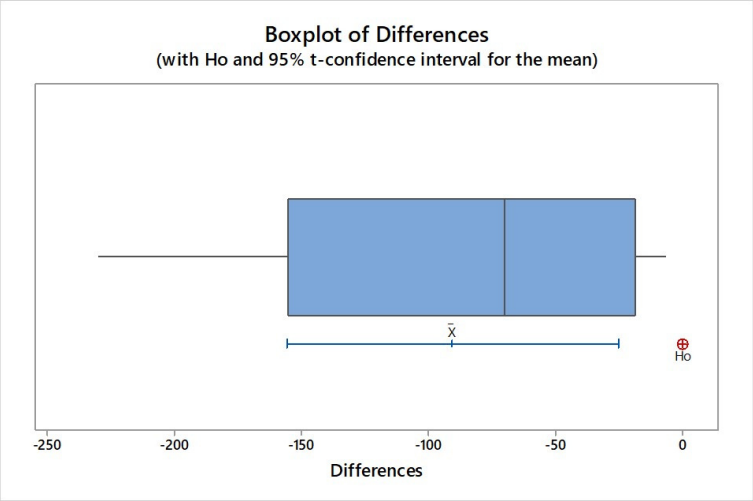

Figure 3: Boxplot for differences between real and simulated activity duration – User 1

equally ($User9$ and 10) in both simulated and real environments to establish whether they are experienced in the use of simulation tools. On the other hand, it was noticed that the activity duration derived from the real environment was significantly higher compared to the simulated data for all the cases in which the null hypothesis was rejected.

Variable 2: *Number of events per activity* (NEPA)

Table 4 and Fig. 4 outline the results of the comparative analysis between simulated and real number of events per activity for User 1 as an example. The p-value ($p = 0.141$) indicates that the data provide support for the null hypothesis. That is, the number of events per activity derived from simulated and real-environment are statistically equivalent for User 1. Based on these findings, we conclude, in case of User 1, that there is no statistically significant difference between simulation and real-environment regarding the number of events per activity with a confidence level of 95%.

A summary of the results derived from the comparative analysis for all users can be found in Table 5. Based on statistical tests, it was concluded that in 75% of the users, the null hypothesis was accepted. It can thus be assumed that the number of events per activity derived from simulation and real-environment tend to be statistically equivalent with a confidence level of 95%.

It is worth noting that users who performed differently regarding activity duration, now have been categorized with p-values lower than the significance level α. It seems that these measures are not strongly correlated and the gap perceived may be due to users who do not have prior experience using simulation tools.

The next step will aim to validate these hypothesis through correlation analysis and other statistical tools allowing us to also identify potential sources of variation. This analysis can be replicated in other comparisons to establish whether the dataset derived from the simulation is equivalent to the real-environment and subsequently

improve the performance of classifiers.

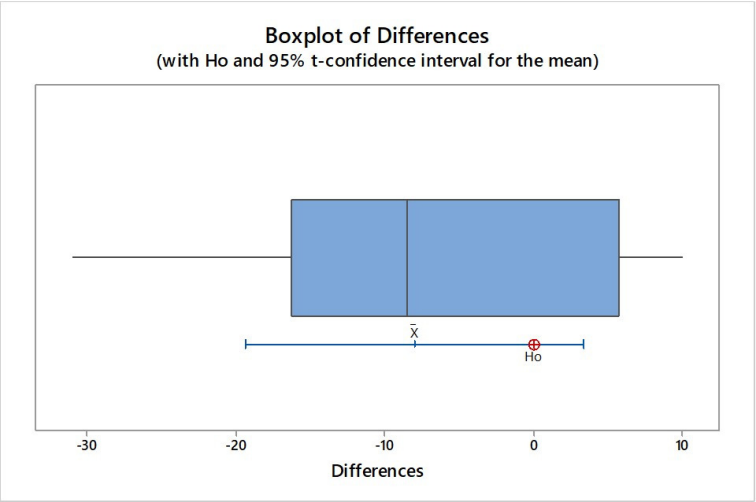

Figure 4: Boxplot for differences between real and simulated number of events per activity – User 1.

Variable 3: *Number of events per type of sensor per activity* (NEPSTPA)

Table 6 and Fig. 4 detail the results of the comparative analysis between the simulated and real number of door sensor events per activity for User 1 as an example. The confidence interval for the mean difference between the numbers of door sensor events per activity values does not include zero, which evidences a significant difference. This is also confirmed by the p-value ($p = 0.014$) that indicates that the data are consistent with the alternative hypothesis. That is, the number of door sensor events ~~per type of sensor (door)~~ per activity, derived from simulated and real-environment, are statistically different for User 1 with a confidence level of 95%. Specifically, this variable is significantly higher in the real-environment.

According to the results provided in Table 7, we found that for 75% of the users the null hypothesis was accepted. It can be therefore assumed that the number of events per door sensor per activity derived from simulation and real-environment are statistically equivalent with a confidence level of 95%.

On the other hand, when considering the pressure sensor, for 75% of users, the null hypothesis was rejected. Thus, the number of events per pressure sensor per activity tend to be statistically different with a confidence level of 95%. In particular, the number of events of the pressure sensor are higher in the real environment.

3.2 Modifying simulated data for predicting real data: The use of regression analysis

Considering the fact that the null hypothesis was rejected in most of the comparisons made between the real and simulated activity duration and the number of events per door sensor per activity, the next question is: *How can simulated data be adjusted in order to better reflect real data?* For this purpose, two types of regression-based

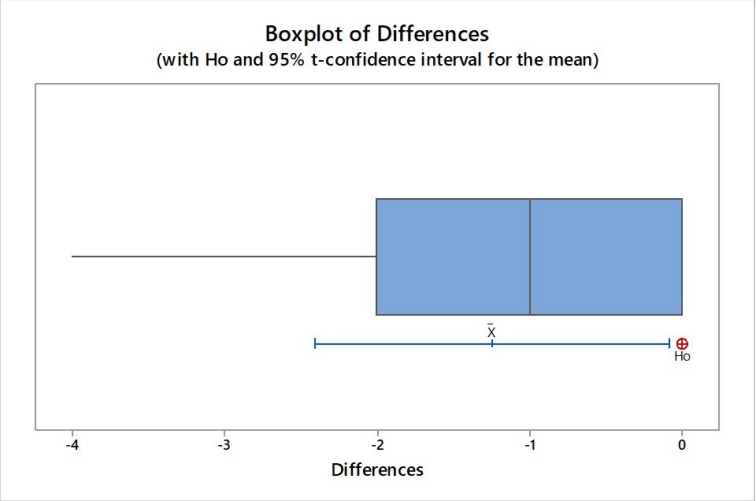

Figure 5: Boxplot for differences between real and simulated number of events per door sensor per activity – User 1.

approaches were investigated: *Activity-based regression models* and *regression model with dummy variables*.

3.2.1 Activity-based regression model

A regression equation was developed for each ADL using Minitab 17® software. The regression assumptions were also validated to determine whether they could be used in practice (see Section 2.2). These models enables the transformation of simulated data to more realistic observations so that they can be used to train activity recognition models. Given that activities are very different from each other, we chose to develop separate models for each considered activity.

- Activity 1: *Go to bed*

  The p-value (0.003) for the regression model indicated in Eq. 7 below showed that the model was significant at a level of 5%. This implies that at least one coefficient is significantly different from zero. The p-values for the estimated coefficients of both *Activity duration* and *Number of events per activity* derived from the simulation were 0.001 and 0.004 (both below the 5% level), and they were therefore correlated to the real activity duration. This suggested that a model with both predictors may be more appropriate.

  The determination coefficient (R^2) told that the predictors explained 95.52% of the total variance in *real activity duration*. The adjusted version (R^2_{adj}) was found to be 92.16%, which demonstrated high fit provided by the model. The predicted determination coefficient, R^2_{pred}, was 85.53%. Since R^2_{pred} was close to the R^2 and R^2_{adj}, the model did not appear to be overfitted and had adequate predictive ability, which was in accordance with[34] where similar situations were considered. Consequently, there were many reasons for assuming the derived regression model developed for the prediction of real activity durations in *Go to bed* as

$$\ln Y = 5.466 - 0.06857X_1 + 0.1026X_2 \tag{7}$$

as an adequate one. Here Y is the response variable `AD_Go to_bed_Real`, X_1 is the covariate `AD_Go to_bed_Sim` and X_2 is the covariate `NEPA_Go to_bed_Sim`. An optimal λ was estimated to be -0.0144583 in order to improve the predictive ability and fit of the regression model. A linear model provided a medium-high performance with determination coefficient $R^2 = 82.53\%$, $R^2_{\text{adj}} = 75.55\%$ and $R^2_{\text{pred}} = 56.99\%$. Therefore, an Euler regression model was explored aiming to achieve better results.

When validating the regression assumptions (normality, homoscedasticity and independence) through the residuals, all of them were found to be satisfied. Particularly, the normality was verified by applying an Anderson-Darling test where $AD = 0.279$, with a p-value of 0.5444 and mean 0. More to the point, for independence validation, the Durbin-Watson statistic $D = 3.2430$ was calculated. In this case ($k' = 2, n = 8$), the lower bound $L = 0.345$ and upper bound $U = 1.489$. As $D > U$, no correlation could be claimed to exist. Finally, unequal variances were not observed and hence, there was no evidence that the spread of residual values tend to increase with increased fitted values.

- Activity 2: *Use bathroom*

  The p-values for the predictors *Activity duration* (0.000) and *Activity duration*^{2} (0.002) were lower than the level of significance $\alpha = 0.05$ and they were hence deemed adequate for a model of the response variable. This was an indication that an expression with these predictors was appropriate. The coefficient of determination, R^2, for the model of *real activity duration* was 97.90%. In addition, the adjusted determination coefficient, R^2_{adj}, was found to be 97.20%, which supports the good fit provided by the model. The prediction performance (R^2_{pred}) for this case was 95.34%. Considering the proximity among R^2, R^2_{adj} and R^2_{pred}, the model was not found to be overfitted and provided high-precision predictions. For the parameter λ the value 0 was used to improve the prediction performance and fit of the regression model. In this case, an Euler regression model was proposed to achieve better results. Consequently, the regression model developed for the prediction of real activity durations for *Use bathroom* is

$$\ln Y \;=\; 0.2006 X_1 - 0.002229 X_1^2 \tag{8}$$

  Here Y is the response variable `AD_Use bathroom` and X_1 is the covariate `AD_Use bathroom_Sim`. In this ADL, a logarithmic regression model was suggested to obtain better results.

  The regression assumptions were verified and found as satisfied through a residual analysis. In particular, the normality was validated using Anderson-Darling test where $AD = 0.446$, the p-value = 0.204 and the mean was equal to 0. For independence verification, the Durbin-Watson statistic $D = 2.976$ was calculated. Considering that ($k' = 1, n = 8$), the lower bound $L = 0.497$ and upper bound $U = 1.003$. As $D > U$, no correlation exists. In this case, unequal variances were not detected and therefore, there was no further evidence that the spread of residual values tend to increase as the fitted values increase.

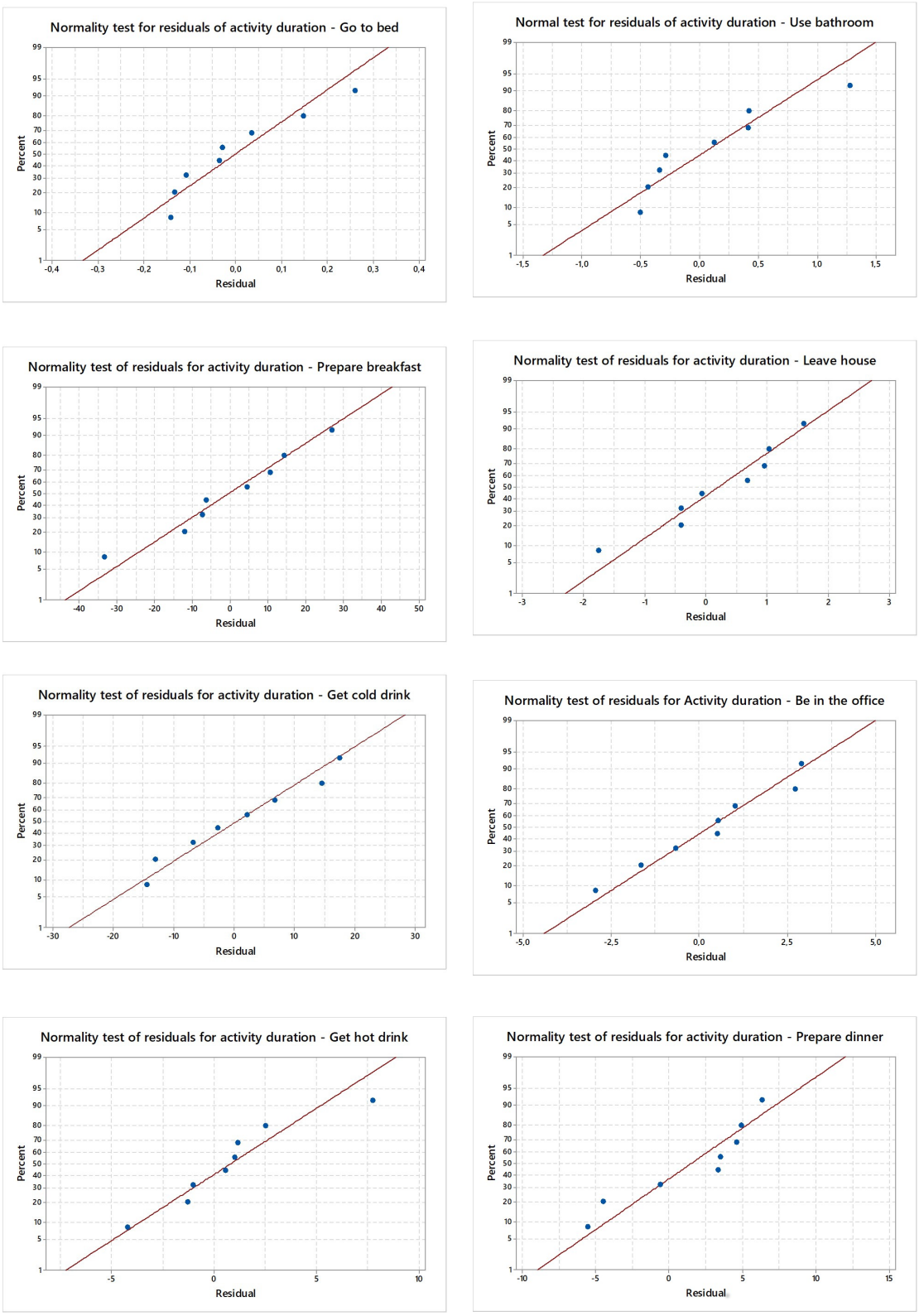

Figure 6: QQ plots of residuals of activity duration. The activities are in the first row from the left: *Go to bed*, *Use the bathroom*, in the second row: *Prepare breakfast*, *Leave house*, in the third row: *Get cold drink*, *Be in office* and in the fourth row: *Get hot drink* and *Prepare dinner*.

- Activity 3: *Prepare breakfast*

  In *Prepare breakfast*, the p-value for the regression model was 0.000 and the regression model was therefore concluded to be significant at the 5% level. Furthermore, the p-values for the predictors *Activity duration* (0.001) and (*Activity duration*)^2 (0.019) were also well below the 5% level and thus significant for the response. These results revealed that a model with these predictors may provide a good performance. The determination coefficient (R^2) was calculated as 96.98% while the R^2_{adj} was found to be 95.97%. Both metrics indicate that the model fits the data well. In addition, $R^2_{\text{pred}} = 92.32\%$. Here R^2, R^2_{pred} and R^2_{adj} were found to be close to each other, and therefore the model was not considered to be overfit and had very good predictive performance. To this end, a quadratic regression model was suggested. The regression model developed for the prediction of real activity durations in *Prepare breakfast* is

$$Y = 3.372X_1 - 0.02722X_1^2 \tag{9}$$

  Here Y is the response variable `AD_Prepare breakfast` and X_1 is the covariate `AD Use bathroom_Sim`. The regression assumptions were then validated and assumed to be satisfied considering the results of the residual analysis. In detail, the normality was assessed by applying an Anderson-Darling test where $AD = 0.179$, p-value = 0.879 and the mean was approximately equal to 0. For the independence validation, the Durbin-Watson statistic D (1.294) was estimated. Considering that ($k' = 1, n = 8$), the lower bound L and upper bound U were established as 0.497 and 1.003 respectively. As $D > U$, no correlation could be discerned. As also seen in previous ADLs, unequal variances were not detected and there was then no evidence of heteroscedasticity.

- Activity 4: *Leave house*

  Regarding *Leave house*, the p-value for the regression model was 0.000, i.e. significant at a level of 5%. Besides, the p-value for the predictor *Activity duration* (0.000) was lower than the 5% level and therefore it was inferred to be significant for the response variable. This predictor was then incorporated with the model to provide better predictive ability and fit.

  Further, the determination coefficient, R^2, was 97.08% while R^2_{adj} was found to be 96.66%. These results suggested a very good fit for the data. Also, R^2_{pred} was found to be equal to 96.23%. Taking into account the proximity among R^2, R^2_{adj} and R^2_{pred}, the model was not overfitted and had superior prediction performance. In this case, a square root regression model (with $\lambda = 0$) was concluded to offer very good results. As a consequence, the regression model provided for the prediction of real activity durations in *Leave house* is

$$Y = 0.0238X_1^2 \tag{10}$$

  Similar to the previous models, Y represents the response variable, in this case `AD_Leave house` and X_1 is the covariate `AD Leave house_Sim`. The normality, homoscedasticity and independence assumptions were also tested and found not to be violated. Specifically, the normality was validated using the

Anderson-Darling test where $AD = 0.212$ and $p - value = 0.779$. On the other hand, the Durbin-Watson statistic D (2.450) was calculated for auto-correlation assumption. Considering that $(k' = 1, n = 8)$, the lower L and upper U were established as 0.497 and 1.003 respectively. As $D > U$, no correlation exists. Also, equal variances of the residuals were found in this analysis.

- Activity 5: *Get cold drink*

  A regression model was found to be significant (with p-value = 0.000) for activity duration of *Get cold drink*. This suggested that at least one coefficient was different from zero. In addition, the p-values for the predictor *Number of events per activity* (0.000) and (*Number of events per activity*)^2 (0.002) were less than the 5% level and therefore significant for the prediction of activity duration. In this case, the R^2 (96.76%) and R^2_{adj} (95.68%), was concluded to explain a high portion of the variance in *real activity duration*. The message from both of these measures was that the model provided a good fit for the data. Further, there are no significant differences among R^2 (91.61%), R^2_{adj} and R^2_{pred} and thus, there was no sign that the model was overfit. The aforementioned results were provided by a quadratic regression model developed for the prediction of activity duration in *Get cold drink* as

$$Y = 12.61X_1 - 0.5332X_1^2 \tag{11}$$

  Here, Y is the response variable `NEPA_Get cold drink` and X_1 is the covariate `NEPA Get cold drink_Sim`. For this model, the regression assumptions were also evaluated for ensuring a high reliability of the prediction in addition to verifying the presence of potential bias. In this particular case, no violation was found. First, the normality was assessed using an Anderson-Darling test where $AD = 0.199$, $p - value = 0.824$ and the mean was approximately equal to 0. The auto-correlation was tested through the Durbin-Watson statistic D (2.462). Considering $(k' = 1, n = 8)$, the lower bound L and the upper bound U were defined as 0.497 and 1.003 correspondingly. As $D > U$, there are no indications of any correlation. Finally, no evidence was found regarding the violation of homoscedasticity assumption.

- Activity 6: *Be in the office*

  For the variable *Be in the office*, a quadratic regression model was found to offer the best predictive ability and fit (p-value = 0.000). This pointed out that at least one coefficient was different from zero. In addition, the p-values for the predictors *Number of events per activity* (0.000) and (*Number of events per activity*)^2 (0.004) were lower than the 5% level and they were hence significant for the activity duration of *Be in the office*. A model including these predictors was therefore suggested. For this model, the determination coefficient R^2 specified that the predictors accounted for 96.50% of the variance in real activity duration of *Be in the office* whilst R^2_{adj} (95.34%) indicated a high explanatory power of the proposed model. In this case, both coefficients contributed to

the good fit provided by the model. In addition, R^2_{pred} (93.94%) was found to be close to both R^2 and R^2_{adj}; thus, the model was not concluded to be overfit and had high prediction performance. The regression model developed for predicting the real activity durations of *Be in the office* is

$$\sqrt{Y} \;=\; 2.526X_1 - 0.1345X_1^2 \tag{12}$$

Here, Y represents the response variable `AD_Be in the office` and X_1 is the covariate `NEPA Be in the office_Sim`. The model assumptions were also validated through the residual analysis and it was found that all of them are satisfied. In particular, the normality was confirmed using an Anderson-Darling test where $AD = 0.212$ and p-value $= 0.779$. As for auto-correlation validation, the Durbin-Watson statistic D was found to be 1.343. Considering that $(k' = 1, n = 8)$, the lower bound L and the upper bound U were calculated as 0.497 and 1.003 correspondingly. As $D > U$, no correlation was distinguished. Finally, the homoscedasticity of the residuals was also verified.

- Activity 7: *Get hot drink*

  The regression model here provided was found to be significant (p-value $=$ 0.000) at a level 5%. Furthermore, the p-value for the predictors *Activity duration* (0.001) and (*Number of events per activity*)^2 (0.008) were below the 5% level. Hence, they were significant for the activity duration of *Get hot drink*. A quadratic regression model including these predictors was concluded to be appropriate. More to the point, the R^2 (94.43%) and R^2_{adj} (92.58%) explained a high proportion of the variance in *real activity duration*. These values pinpointed that the model fitted the data well. Regarding the prediction performance, R^2_{pred} (90.60%) was found to be reasonably close to R^2 and R^2_{adj}. Therefore, the model

$$\sqrt{Y} \;=\; 0.5641X_1 - 0.1345X_1^2 \tag{13}$$

  was not overfitted and had a high-precision predictive performance (in accordance with[34]).

  In Eq. 13, Y represents the response variable `AD_Get hot drink` and X_1 is the covariate `Get hot drink_Sim`. A quadratic regression model was also found to provide the highest predictive ability and fit. When validating the regression assumptions through the residuals, it was proved that all of them were satisfied. In particular, the normality was checked by means of an Anderson-Darling test where $AD = 0.361$ and p-value $= 0.348$. The Durbin-Watson statistic D (2.656) was estimated to verify the independence assumption. Given that $(k' = 1, n = 8)$, the lower bound L and upper bound U were established as 0.497 and 1.003 respectively. As $D > U$, there are no signs of correlation. Finally, there was also evidence that the homoscedasticity of the residuals is not violated.

- Activity 8: *Prepare dinner*

  For *Prepare dinner*, a square root (with $\lambda = 0.5$) regression model (p-value $=$

0.000) was concluded to provide the highest fit and predictive ability. Indeed, this indicates that at least one coefficient is non-zero. Additionally, the p-value for the predictor *Activity duration* (0.000) was less than the 5% level. It was therefore significant for the activity duration of *Prepare dinner* and should be included in the prediction model. The determination coefficient, R^2, for this model was 85.75% whilst R^2_{adj} (83.71%) confirmed a good fit provided by the model. In addition, R^2_{pred} (80.12%) was close to the R^2 and adjusted R^2 values. Based on these results, the model

$$Y = 0.053X_1^2 \tag{14}$$

was not concluded to be overfit and had an acceptable predictive capability.

In this case, Y denotes the response variable `AD_Prepare dinner` and X_1 represents the covariate `AD Prepare dinner_Sim`. In this case, a quadratic regression model was found to provide the highest predictive performance and fit. Similar to the above mentioned ADLs, the residual analysis supported the regression assumptions. More precise, the normality was tested with the Anderson-Darling statistic where $AD = 0.526$, p-value $= 0.121$ and the mean was approximately equal to 0. To validate the independence of residuals, the Durbin-Watson statistic D (2.003) was calculated. Considering that $(k' = 1, n = 8)$, the lower bound L and the upper bound U were established as 0.497 and 1.003 respectively. As $D > U$, no correlation could be concluded. Finally, no evidence was found for rejecting the homoscedasticity of residuals.

Please, refer to Table 9 for a summarized presentation of the R^2 values for the regression analyses of the different activities. Table 9 also contains the validation results for *normality*, *independence* and *homoscedascticity* assumptions.

3.2.2 Regression model with dummy variables

A general regression model with *dummy variables* was also explored. These variables act as *switches* turning several parameters on and off in the predictive equation. In this case, they represent the type of ADL and assume the value of 0 or 1 indicating the presence or absence of a particular ADL. The *dummy variables* are defined as follows:

D_1 (Go to bed): $D_1 = 1$ if the ADL is *Go to bed*, 0 otherwise.

D_2 (Use bathroom): $D_2 = 1$ if the ADL is *Use bathroom*, 0 otherwise.
D_3 (Prepare breakfast): $D_3 = 1$ if the ADL is *Prepare breakfast*, 0 otherwise.
D_4 (Leave house): $D_4 = 1$ if the ADL is *Leave house*, 0 otherwise.
D_5 (Get cold drink): $D_5 = 1$ if the ADL is *Get cold drink*, 0 otherwise.
D_6 (Be in the office): $D_6 = 1$ if the ADL is *Be in the office*, 0 otherwise.
D_7 (Get hot drink): $D_7 = 1$ if the ADL is *Get hot drink*, 0 otherwise.
D_8 (Prepare dinner): $D_8 = 1$ if the ADL is *Prepare dinner*, 0 otherwise.

In addition, X_1 (*Activity duration*) and X_2 (*Number of events per activity*) were included in the predictive model. Table 8 describes the set of predictors that were found to be significant for real activity duration Y at a significance level of 5%. This suggests that a model with these predictors may be more suitable.

In Table 2, R^2 told that the significant predictors explained 98.69% of the variance in real activity duration Y. The value of the adjusted determination coefficient R^2_{adj} was found to be 98.48%, supporting an excellent fit. The predicted determination coefficient R^2_{pred} was 97.03%), close to the R^2 and R^2_{adj} values. Hence, the model was not overfitted (in accordance with[34]).

Also, $\lambda = 0$ proved to improve the predictive ability and fit of the regression model. In this case, the square root regression model

$$\begin{aligned} \ln Y \;\; = \;\; & 0.175X_2 + 0.761D_1 + 2.780D_2 - 0.004\,X_1{*}X_2 - 0.004\,X_1{*}D_2 + \\ & + 0.009\,X_1{*}D_3 + 0.037\,X_1{*}D_4 + 0.017\,X_1{*}D_5 + 0.00002\,X_1^2{*}X_2 \end{aligned} \tag{15}$$

was concluded to offer very good results as seen below in Figures 7 and 8.

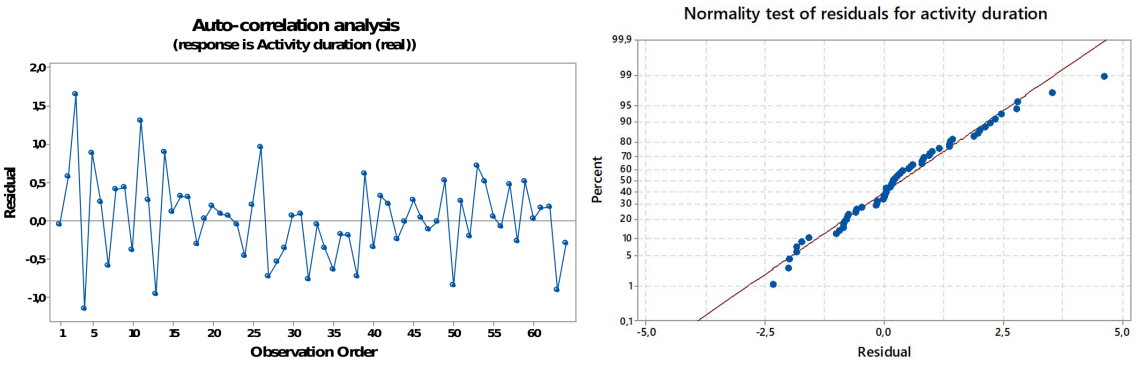

Figure 7: Auto-correlation and QQ-plot of activity duration.

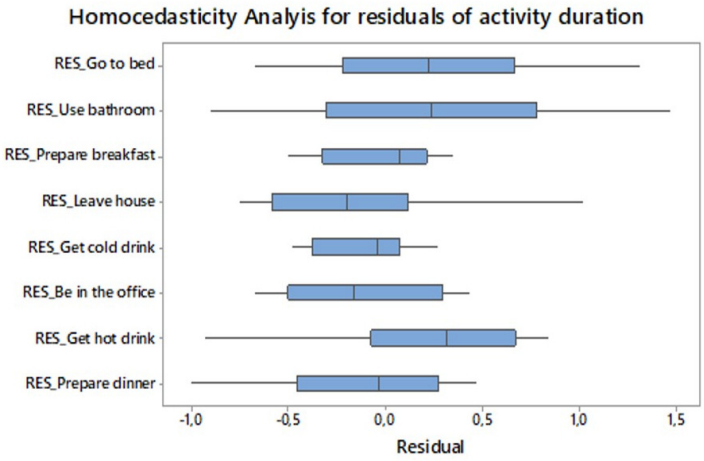

Figure 8: Homoscedasticity for residuals of activity duration.

When validating the regression assumptions through the residuals, all of them were found to be satisfied. In particular, normality was verified by applying an Anderson-Darling test resulting in $AD = 0.427$, p-value $= 0.304$ and mean approximately equal to zero. When checking independence, the Durbin-Watson statistic D was equal to 2.3502. In this case ($k' = 9, n = 64$), the lower bound L and the upper bound U were established as 1.1084 and 1.771 respectively. As $D > U$, there were no signs of correlation. Finally, unequal variances were not concluded using Bartlett method (p-value $= 0.167$) and there was no evidence that the spread of residual values tend to increase as the fitted values increase.

4 Conclusions

Caring for older adults living alone can be made safer and less costly by automatically monitoring how they perform ~~activities of daily living (~~ADLs~~)~~ using smart home sensors. One important step in this process is the automatic detection and recognition of which activity is being performed. Accurate activity recognition models are highly dependent on the availability of adequate and sufficient data. Unfortunately, the acquisition of large amounts of data is costly and resource intensive.

We postulate that a more cost-effective solution to acquiring data is the use of simulation tools and synthetic datasets. The main challenge with this approach is the generation of synthetic data with the exact same characteristics as real data. It is important to note that simulated data can be considerably different from real data and may depend on the experience of the simulation operator. In this regard, significant differences were found regarding the activity duration and the number of events per door sensor.

In this work we evaluate the characteristics of simulated data sets with respect to real data, and propose the use of regression models to transform simulated data in order to better represent real observations. All single activity models were subject to simple regression (i.e. a single covariate) with one exception: activity *1. Go to bed* which included both covariates *Activity Duration* and *Number of Events Per Activity*. It turns out that the activities *1. Go to bed* and *2. Use bathroom* (see Equations 7 and 8) were successfully modeled by variants of the logarithmic transform as defined in Equation 3. Further, the activities *3. Prepare breakfast*, *4. Leave home*, *5. Get cold drink* and *8. Prepare dinner* (see Equations 9, 10, 11 and 14) are well captured by quadratic transform models as defined in Equation 4 but without intercept and regarding activities 4. and 8. just with the quadratic term. Activities *6. Be in office* and *7. Hot drink* (see Equations 12 and 13) were modeled by squared quadratic transform as defined in Equation 5. All covariates were included in an omnibus model (see Equation 15) including both dummy variables and interaction terms. Results demonstrate that simulated data can be post-processed to better approximate real data ($R^2_{\text{pred}} = 97.03\%$) when using a regression incorporating dummy variables.

We have, in this work, only considered the duration and intensity of sensor activations, regardless of sensor types. However, human behaviour captured as a sequence of sensor events is in general complex and may (besides duration and number of events) contain permutations of events within a sequence or for the subsequences of a sequence. An interesting direction to investigate is how to assess the realism of synthetic data given the statistical properties of the sensor event ordering. Future work will consider the different types of sensors involved in each activity so as to improve the accuracy of transformations.

Acknowledgments

The Authors which to acknowledge support from the REMIND Project from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734355.

References

[1] Organization World Health. *World report on ageing and health.* World Health Organization 2015.

[2] DESA UN. United nations department of economic and social affairs, population division. world population prospects: The 2015 revision, key findings and advance tables in *Technical Report*Working Paper No. ESA/P/WP. 241 2015.

[3] Prince MJ, Wu F, Guo Y, et al. The burden of disease in older people and implications for health policy and practice *The Lancet.* 2015;385:549–562.

[4] Paterson C. *World Alzheimer Report 2018.* Alzheimer's Disease International 2018.

[5] Stepler R. *Smaller share of women ages 65 and older are living alone: More are living with spouse or children.* Pew Research Center 2016.

[6] National Statistics Office. Labour force survey 2018.

[7] Barrios Miguel Ortiz, Jiménez Heriberto Felizzola. Reduction of average lead time in outpatient service of obstetrics through six sigma methodology in *Ambient Intelligence for Health*:293–302Springer 2015.

[8] Holmes J. An overview of the domiciliary care market in the UK 2016.

[9] Alberdie Ane, Weakley Alyssa, Schmitter-Edgecombe Maureen, et al. Smart Home-Based Prediction of Multidomain Symptoms Related to Alzheimer's Disease *IEEE Journal of Biomedical and Health Informatics.* 2018;22:1720–1731.

[10] Mlinac ME, Feng MC. Assessment of activities of daily living, self-care, and independence *Archives of Clinical Neuropsychology.* 2016;31:506–516.

[11] Millan-Calenti JC, Tubío J, Pita-Fernández S, et al. Prevalence of functional disability in activities of daily living (ADL), instrumental activities of daily living (IADL) and associated factors, as predictors of morbidity and mortality *Archives of gerontology and geriatrics.* 2010;50:306–310.

[12] Debes C, Merentitis A, Sukhanov S, Niessen M, Frangiadakis N, Bauer A. Monitoring activities of daily living in smart homes: Understanding human behavior *IEEE Signal Processing Magazine.* 2016;33:81–94.

[13] De-La-Hoz-Franco Emiro, Ariza-Colpas Paola, Quero Javier Medina, Espinilla Macarena. Sensor-Based Datasets for Human Activity Recognition–A Systematic Review of Literature *IEEE Access.* 2018;6:59192–59210.

[14] Alshammari N, Alshammari T, Sedky M, Champion J, Bauer C. OpenSHS: Open Smart Home Simulator *Sensors.* 2017;17:1003.

[15] Krishnan NC, Cook DJ. Activity recognition on streaming sensor data *Pervasive and mobile computing.* 2014;10:138–154.

[16] Helal S, Lee JW, Hossain S, Kim E, Hagras H, Cook D. Persim-Simulator for human activities in pervasive spaces in *Intelligent Environments (IE), 2011 7th International Conference on*:192–199IE 2011.

[17] Synnott J, Nugent C, Jeffers P. Simulation of smart home activity datasets *Sensors.* 2015;15:14162–14179.

[18] Mendoza-Palechor Fabio, Menezes Maria Luiza, Sant'Anna Anita, Ortiz-Barrios Miguel, Samara Anas, Galway Leo. Affective recognition from EEG signals: an integrated data-mining approach *Journal of Ambient Intelligence and Humanized Computing.* 2018:1–20.

[19] Hamad R, Järpe E, Lundström J. Stability Analysis of the t-SNE Algorithm for Human Activity Pattern Data in *The 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC2018)* 2018.

[20] Helal S, Kim E, Hossain S. Scalable approaches to activity recognition research in *Proceedings of the 8th International Conference Pervasive Workshop*:450–453 2010.

[21] Ortiz Miguel, Neira Dionicio, Jiménez Genett, Hernández Hugo. Solving flexible job-shop scheduling problem with transfer batches, setup times and multiple resources in apparel industry in *International Conference in Swarm Intelligence*:47–58Springer 2016.

[22] Alshammari T, Alshammari N, Sedky M, Howard C. SIMADL: Simulated Activities of Daily Living Dataset *Data.* 2018;3:11.

[23] Francillette Y, Boucher E, Bouzouane A, Gaboury S. The Virtual Environment for Rapid Prototyping of the Intelligent Environment *Sensors.* 2017;17:2562.

[24] Lee JW, Cho S, Liu S, Cho K, Helal S. Persim 3d: Context-driven simulation and modeling of human activities in smart spaces *IEEE Transactions on Automation Science and Engineering.* 2015;12:1243–1256.

[25] Kamara-Esteban O, Azkune G, Pijoan A, Borges CE, Alonso-Vicario A, Ipiña D López. MASSHA: an agent-based approach for human activity simulation in intelligent environments *Pervasive and Mobile Computing.* 2017;40:279–300.

[26] Synnott J, Nugent C, Zhang S, et al. Environment simulation for the promotion of the open data initiative in *Smart Computing (SMARTCOMP), 2016 IEEE International Conference on*:1–6IEEE 2016.

[27] Larsen RJ, Marx ML. *An Introduction to Mathematical Statistics and its Applications.* Pearson6 ed. 2006.

[28] Vittinghoff E, Gidden DV, Shiboski SC. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models.* Springer2 ed. 2011.

[29] Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press 2006.

[30] Suits DB. Use of Dummy Variables in Regression Equations *Journal of the American Statistical Association.* 1957;52:548–551.

[31] Gergonne JD. The application of the method of least squares to the interpolation of sequences *Historia Mathematica.* 1974;1:439–437. Translated by Ralph St. John and Stephen M. Stigler from the 1815 French edition.

[32] Lundström J, Morais WO De, Menezes M, et al. Halmstad Intelligent Home-Capabilities and Opportunities in *International Conference on IoT Technologies for HealthCare*:9–15Springer 2016.

[33] Nugent Chris, Synnott Jonathan, Gabrielli Celeste, et al. Improving the quality of user generated data sets for activity recognition in *Ubiquitous Computing and Ambient Intelligence*:104–110Springer 2016.

[34] Statistics Minitab 2003. http://www.scribd.com/document/98819705/5-StatisticsAllTopics.

Tables

Variable	*Mean*	*Standard dev.*	*S.E. of the mean*
AD_User 1_Simulation	46.5	16.5	5.8
AD_User 1_Real environment	137.1	77.1	27.3
Difference	−90.6	77.8	27.5

Table 1: Paired t-test results for comparison between real and simulated activity duration in User1.

S	R^2	95% C.I.	R^2 (*adj.*)	95% C.I.	R^2 (*pred*)	95% C.I.	*PRESS*
0.5515	0.9869	[0.7843, 1]	0.9848	[0.7827; 1]	0.9703	[0.6308; 1]	38.0572

Table 2: Summary of determination coefficient values for the regresion model based on dummy variables.

User code	*Confidence interval for the difference in seconds (95%)*	*t-value*	*p-value*	*Conclusion*
001	[−155.7, −25.5]	−3.29	0.013	Statistically different
002	[−68.6, −7.4]	−2.93	0.022	Statistically different
003	[−121.3, −22.0]	−3.41	0.011	Statistically different
004	[−79.2, −5.3]	−2.70	0.031	Statistically different
005	[−170.0, −10.5]	−2.68	0.032	Statistically different
006	[−132.4, −14.6]	−2.95	0.021	Statistically different
007	[−63.1, 41.8]	−0.48	0.647	Statistically equivalent
008	[−68.8, 15.0]	−1.52	0.173	Statistically equivalent

Table 3: Results of comparative analysis between simulated and real data in terms of Activity duration

Variable	*Mean*	*Standard dev.*	*S.E. of the mean*
NEPA_User1_Simulation	10	6.00	2.12
NEPA_User1_Real environment	18	8.96	3.17
Difference	−8	13.63	4.82

Table 4: Paired t-test results for comparison between real and simulated number of events per activity in User 1.

User	*95% C.I. for the difference*	*t-value*	*p-value*	*Conclusion*
001	$[-19.39, 3.39]$	-1.66	0.141	Statistically equivalent
002	$[-3.22, 4.47]$	0.38	0.712	Statistically equivalent
003	$[-11.90, 2.65]$	-2.91	0.023	Statistically different
004	$[-11.90, 2.65]$	-1.50	0.176	Statistically equivalent
005	$[-13.60, 3.10]$	-1.49	0.180	Statistically equivalent
006	$[-159.1, -33.4]$	-3.62	0.009	Statistically different
007	$[-9.52, 20.02]$	0.84	0.428	Statistically equivalent
008	$[-12.29, 24.29]$	0.78	0.463	Statistically equivalent

Table 5: Results of comparative analysis between simulated and real data in terms of Number of events per activity

Variable	*Mean*	*Standard dev.*	*S.E. of the mean*
NEPSTPA_DOOR_Simulation	6.75	3.96	1.40
NEPSTPA_DOOR_Real environment	8.13	4.29	1.52
Difference	-1.38	1.19	0.42

Table 6: Paired t-test results for comparison between real and simulated number of events per DOOR sensor per activity.

User	*Type of sensor*	*95% C.I. for the difference*	*t-value*	*p-value*	*Conclusion*
001	DOOR	$[-2.368. - 0.382]$	-3.27	0.014	Different
	PRESSURE	$[-11.62, -3.63$	-4.51	0.003	Different
002	DOOR	$[-3.043, 0.293]$	-1.95	0.092	Equivalent
	PRESSURE	$[-7.68, -0.07]$	-2.41	0.047	Different
003	DOOR	$[-3.095, 0.845]$	-1.35	0.219	Equivalent
	PRESSURE	$[-8.95, -1.30]$	-3.16	0.016	Different
004	DOOR	$[-3.095, 0.845]$	-1.35	0.219	Equivalent
	PRESSURE	$[-8.52, -1.23]$	-3.16	0.016	Different
005	DOOR	$[-2.715, 0.215]$	-2.02	0.083	Equivalent
	PRESSURE	$[-11.69, -2.81]$	-3.86	0.006	Different
006	DOOR	$[-3.248, -0.502]$	-3.23	0.014	Different
	PRESSURE	$[-8.99, 0.74]$	-2.01	0.085	Equivalent
007	DOOR	$[-6.52, 0.52]$	-2.02	0.084	Equivalent
	PRESSURE	$[-6.60, 2.10]$	-1.22	0.261	Equivalent
008	DOOR	$[-3.670, 0.420]$	-1.88	0.102	Equivalent
	PRESSURE	$[-6.34, -1.41]$	-3.72	0.007	Different

Table 7: Results of comparative analysis between simulated and real data in terms of Number of events per type of sensor per activity.

Predictor	*DF*	*Seq SS*	*Contribution*	*Adj SS*	*Adj MS*	*F-value*	*P-value*
X_2	1	743.37	0.5804	5.02	5.018	16.49	0.000
D_1	1	431.09	0.3366	4.98	4.979	16.37	0.000
D_2	1	44.45	0.0347	23.42	23.421	76.99	0.000
$X_1 * X_2$	1	11.05	0.0086	4.16	4.157	13.66	0.001
$X_1 * D_2$	1	0.03	0.0001	1.70	1.699	5.58	0.022
$X_1 * D_3$	1	2.25	0.0018	9.58	9.583	31.50	0.000
$X_1 * D_4$	1	24.96	0.0195	21.94	21.944	72.13	0.000
$X_1 * D_5$	1	3.59	0.0028	5.30	5.299	17.42	0.000
$X_1^2 * X_2$	1	3.26	0.0025	3.26	3.260	10.72	0.002
Error	55	16.73	0.0131	16.73	0.304		
Total	64	1280.79	1				

Table 8: ANOVA analysis for the regression model with dummy variables.

ADL	*Go to bed*	*Use bathroom*	*Prepare breakfast*	*Leave house*	*Get cold drink*	*Be in office*	*Get hot drink*	*Prepare dinner*
Regression model	Log.	Log.	Quadr.	Squared quadr.	Quadr.	Quadr.	Quadr.	Squared quadr.
R^2	0.9042	0.9790	0.9698	0.9708	0.9676	0.9650	0.9443	0.8575
R^2 (*adj.*)	0.8659	0.9720	0.9597	0.9666	0.9568	0.9534	0.9258	0.8371
R^2 (*pred.*)	0.7002	0.9534	0.9232	0.9623	0.9161	0.9394	0.9060	0.8012
Residual analysis								
Durbin-Watson	2.737	2.976	1.294	2.450	2.463	1.344	2.657	2.004
Normality p-value	0.325	0.204	0.879	0.779	0.824	0.779	0.348	0.121
Homoscedasticity p-value	0.303	0.445	0.445	0.127	0.445	0.537	0.445	0.537

Table 9: Summary of determination coefficient values for the different activities.